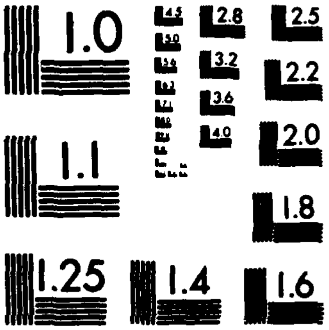
AD-A186 033   STRATEGIES OF DATA ANALYSIS(U) PITTSBURGH UNIV PA   1/1
CENTER FOR MULTIVARIATE ANALYSIS C R RAO JUN 87
TR-87-14 AFOSR-TR-87-1088 F49620-85-C-0008

UNCLASSIFIED   F/G 12/3   NL



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

①

AFOSR-TR- 87-1088

STRATEGIES OF DATA ANALYSIS*

BY

C. Radhakrishna Rao

Center for Multivariate Analysis
University of Pittsburgh

Center for Multivariate Analysis

University of Pittsburgh

1

STRATEGIES OF DATA ANALYSIS*

BY

C. Radhakrishna Rao

Center for Multivariate Analysis

University of Pittsburgh

June 1987

Technical Report No. 87-14

Center for Multivariate Analysis
Fifth Floor, Thackeray Hall
University of Pittsburgh
Pittsburgh, PA 15260

*Research sponsored by the Air Force Office of Scientific Research (AFSC) under contract F49620-85-C-0008. The United States Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation hereon.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER AFOSR-TR- 87-1088 ~~-87-14~~	2 GOVT ACCESSION NO. ADA186033	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Strategies of Data Analysis		5. TYPE OF REPORT & PERIOD COVERED Journal ~~technical~~ - June 1987
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) C. Radhakrishna Rao		8. CONTRACT OR GRANT NUMBER(s) F49620-85-C-0008
9. PERFORMING ORGANIZATION NAME AND ADDRESS Center for Multivariate Analysis University of Pittsburgh, 515 Thackeray Hall Pittsburgh, PA 15260		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 61102F 2304 A5
11. CONTROLLING OFFICE NAME AND ADDRESS Air Force Office of Scientific Research Department of the Air Force Bolling Air Force Base, DC 20332 NM		12. REPORT DATE June 1987
		13. NUMBER OF PAGES 18
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) Same as 11		15. SECURITY CLASS. (of this report) unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19 KEY WORDS (Continue on reverse side if necessary and identify by block number)

Cross-examination, cross-validation, exploratory data analysis, graphical techniques, inferential data analysis, jack-knife, non-response, outliers, robustness, specification.

20 ABSTRACT (Continue on reverse side if necessary and identify by block number)

The purpose of statisitcal analysis is "to extract all the information from observed data". The recorded data may have some defects such as recording errors and outliers and the first task of a statistician is to scrutinize or corss-examine the data for possible defects and understand its special features. The next step is the specification of a suitable stochastic model for the data using prior information and cross-validation techniques. On the basis of a chosen model, inferential analysis is made, which comprises of estimation of unknown parameters, tests of hypotheses, prediction of future observations an ddecision making. Examing data under different possible models is suggested as more informative than using robust procedures to safeguard against possible alternative models. Finally data analysis must also provide information for raising new questions and planning future investigations. Some aspects of data analysis as outlined above are illustrated through examples.

STRATEGIES OF DATA ANALYSIS

C. Radhakrishna Rao
Center for Multivariate Analysis
University of Pittsburgh, Pittsburgh, PA 15260

1. Historical

Styles in statistical analysis change over time while the object of "extracting all the information from data" or "summarization and exposure" remains the same. Statistics has not yet aged into a stable discipline with complete agreement on foundations. Certain methods become popular at one time and are replaced in course of time by others which look more fashionable. In spite of the controversies, the statistical methodology and fields of applications are expanding. The computer and more specifically the availability of graphic facilities have had a greater impact on data analysis. It may be of interest to briefly review some historical developments in data analysis.

It has been customary to consider descriptive and theoretical statistics as two branches of statistics with distinct methodologies. In the former, the object is to summarize a given data set in terms of certain "descriptive statistics" such as measures of location and dispersion, higher order moments and indices, and also to exhibit salient features of the data through graphs such as histograms, bar diagrams and two dimensional charts. No reference is made to the stochastic mechanism (or probability distribution) which gave rise to the observed data. The descriptive statistics thus computed are used to compare different data sets. Even some rules are provided for the choice among alternative statistics, such as the mean, median and mode, depending on the nature of the data set. Such statistical analysis is referred to as descriptive data analysis (DDA). In theoretical statistics, the object is again summarization of data, but with reference to a specified family of underlying probability distributions. The summary or descriptive statistics in such a case heavily depend on the specified stochastic model and their distributions are used to specify margins of uncertainty in inference about the unknown parameters. Such methodology is referred to as inferential data analysis (IDA).

Karl Pearson (K.P.) was the first to try to bridge the gap between DDA and IDA. He used the insight provided by the descriptive analysis based on moments and histograms to draw inference on the underlying family of distributions. For this purpose he invented the first and perhaps the most important test criterion, the chi-squared statistic, to test the hypothesis that the given data arose from a specified probability distribution, which "ushered in a new sort of decision making" [See Hacking (1984), where K.P.'s chi-squared is eulogized as one of top 20 discoveries since 1900. Even R.A. Fisher (R.A.F.) expressed his appreciation of K.P.'s chi-squared test in personal conversation with the author.] K.P. also created a variety of probability distributions distinguishable by four moments. A beautiful piece of research work done by K.P. through the use of histograms and chi-squared test is the discovery that the distribution of the size of bacteria found in a certain organism is a mixture of two normal distributions (see Pearson (1948)).

The need to develop general methods of estimation arose in extending the chi-squared test to examine a composite hypothesis that the underlying distribution belongs to a specified parameteric family of distributions. K.P. proposed the estimation of parameters by moments and using the chi-squared test based on the fitted distribution. Certain refinements were made by R.A.F both in terms of obtaining a better fit to given data through the estimation of unknown parameters by the method of maximum likelihood and also in the exact use of the chi-squared test using the concept of degrees of freedom when the unknown parameters are estimated.

During the twenties and thirties, R.A.F. created an extraordinarily rich array of statistical ideas. In a fundamental paper in 1922 he laid the foundations of "theoretical statistics", of analysing data through a specified stochastic model. He developed exact small sample tests for a variety of hypotheses under normality assumption and advocated their use with the help of tables of certain critical values, usually 5% and 1% quantiles of the test criterion. During this period, under the influence of R.A.F., great emphasis was laid on tests of significance and numerous contributions were made by Hotelling, Bose, Roy and Wilks among others to exact sampling theory. Although R.A.F. mentioned specification, the problem first considered by K.P., as an important aspect of statistics in his 1922 paper, he did not pursue the problem further. Perhaps in the context of small data sets arising in biological research which R.A.F. was examining, there was not much scope for investigating the problem of specification or subjecting observed data to detailed descriptive analysis to look for special features or to empirically determine suitable transformations of data to conform to an assumed stochastic model. R.A.F. used his own experience and external information of how data are obtained in deciding on specification. [See the classical paper by R.A.F. (1934) on the effect of methods of ascertainment on the estimation of frequencies.] At this stage of statistical developments inspired by R.A.F.'s approach, attempts were made by others to look for what are called non-parametric test criteria whose distributions are independent of the underlying stochastic model for the data (Pitman (1937)) and to investigate robustness of test criteria proposed by R.A.F. for departures from normality of the underlying distribution.

The twenties and thirties also saw systematic developments in data collection through design of experiments introduced by R.A.F., which enabled data to be analysed in a specified manner through analysis of variance and interpreted in a meaningful way; design dictated the analysis and analysis revealed the design.

While much of the research in statistics in the early stages was motivated by problems arising in biology, parallel developments were taking place in a small scale on the use of statistics in industrial production. Shewhart (1931) introduced simple graphical procedures through control charts for detecting changes in a production process, which is probably the first contribution to detection of outliers or change points.

Much of the methodology proposed by R.A.F. was based on intuition, and no systematic theory of statistical inference was available. This was supplied by J. Neyman and E.S. Pearson in 1928 (see collected papers in 1966) by providing some kind of axiomatic set up for deriving appropriate statistical methods, specially in testing of hypotheses, which was further pursued and perfected by Wald (1950) as a theory for decision making. R.A.F. maintained that his methodology was more appropriate in scientific inference while conceding that the ideas of Neyman and Wald might be more revelant in technological applications, although the latter claimed universal validity for their theories. Wald also introduced sequential methods for application in sampling inspection, which R.A.F. thought has applications in biology also.

The forties saw the development of sample surveys which involved collection of vast amounts of data by investigators by eliciting information from randomly chosen individuals on a set of questions. In such a situation, problems such as ensuring accuracy (free from bias, recording and response errors) and comparability (between investigators and methods of enquiry) of data assumed paramount importance. Mahalanobis (1931, 1944) was perhaps the first to recognize that such errors in survey work are inevitable and could be more serious than sampling errors, and steps should be taken to control and detect these errors in designing a survey and to develop suitable scrutiny programs for detecting gross errors (outliers) and inconsistent values in collected data.

We have briefly discussed what is commonly believed to be two branches of statistics, viz., descriptive and inferential statistics, and the need felt by practicing statisticians to clean the data of possible defects which may vitiate inferences drawn from statistical analysis. What was perhaps needed is an integrated approach, providing a proper understanding of the data, its defects and special features, methods for selection of a suitable stochastic model or a class of models for analysis of data to answer specific questions and to raise new questions for further investigation. A great step in this direction was made by Tukey (1962, 1973) in developing what is known as exploratory data analysis (EDA). The basic philosophy of EDA is to understand the special features of data and to use robust inference procedures to accomodate for a wide class of possible stochastic models for the data. Instead of asking the Fisherian question as to what summary statistics are appropriate for a specified stochastic model, Tukey proposed asking for what class of stochastic models, a given summary statistic is appropriate.

In the present paper, I propose to discuss strategies of data analysis by giving some examples. The scheme of data analysis proposed is exhibited in Chart 1, which is based on my own experience in analysing large data sets and which seems to combine K.P.'s descriptive, Fisher's inferential and Tukey's exploratory data analyses and Mahalanobis' concern for non-sampling errors.

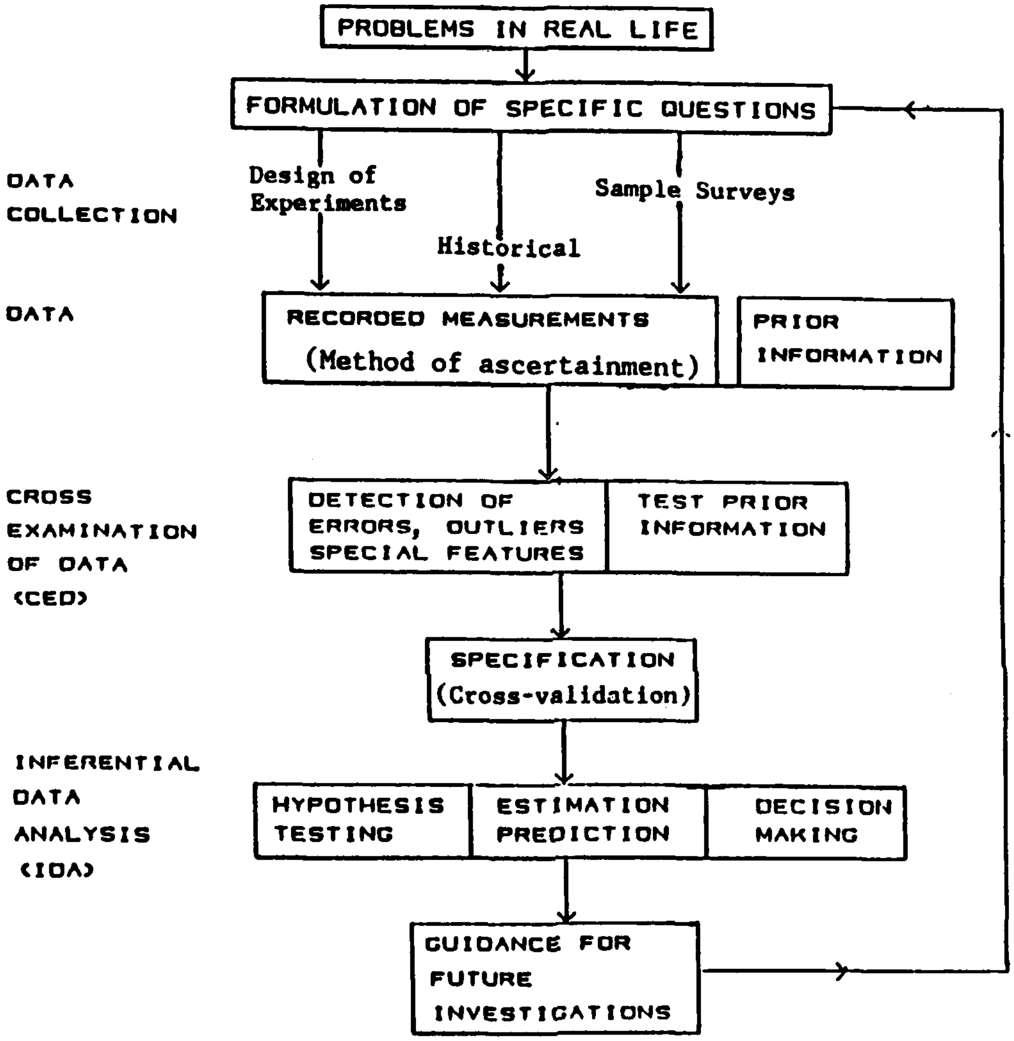
In Chart 1, data is used to represent the entire set of recorded measurements (or observations) and how they are obtained, by an experiment, sample survey or from historical records, and the operational procedures involved in recording the observations, and any prior information on the nature of data or the stochastic model underlying the data.

Cross-examination of data (CED) represents whatever analysis is done to understand the nature of the data, to find measurement and recording errors, to detect outliers, to test validity of prior information and to examine whether data are genuine or faked. The analysis is also intended to select a suitable stochastic model or a class of stochastic models for further analysis of data.

Inferential data analysis stands for the entire body of statistical methods for estimation, prediction, testing of hypothesis and decision making based on a specified stochastic model for observed data. The aim of data analysis should be to extract all available information from data and not merely confined to answering specific questions. Data often contain valuable information to indicate new lines of research and to make improvements in designing future experiments or sample surveys for data collection.

The sequence of data analysis indicated in Chart 1 as CED and IDA should not be regarded as distinct categories with different methodologies. It only shows what we should do to begin with when presented with data and in what form the final results are expressed and used in practical applications. Some results of IDA may suggest further CED, which in turn may indicate changes in IDA.

STRATEGIES OF DATA ANALYSIS


PROBLEMS IN REAL LIFE
FORMULATION OF SPECIFIC QUESTIONS
DATA COLLECTION
Design of Experiments
Historical
Sample Surveys
DATA
RECORDED MEASUREMENTS (Method of ascertainment)
PRIOR INFORMATION
CROSS EXAMINATION OF DATA (CED)
DETECTION OF ERRORS, OUTLIERS SPECIAL FEATURES
TEST PRIOR INFORMATION
SPECIFICATION (Cross-validation)
INFERENTIAL DATA ANALYSIS (IDA)
HYPOTHESIS TESTING
ESTIMATION PREDICTION
DECISION MAKING
GUIDANCE FOR FUTURE INVESTIGATIONS

2. Cross-examination of data

Statisticians are often required to work on data collected by others. Such data may contain recording errors, inconsistent or even faked values and outliers. There is also the possibility that the data were edited and some observations were discarded and not put on record at the discretion of the observer. Some relevant factors for identification and classification of sampled units might not have been recorded leading to clustering of data. Such defects in data would vitiate inferential data analysis unless they are taken into account and suitable modifications are made in data analysis. The process of examining the data for such defects and special features or cross-examination of data, which is "the first task of a statistician" as Fisher put it, is not a routine matter although graphical representation of data through histograms, two dimensional scatter plots and probability plots, and the computation of certain descriptive statistics would be of great help. Much depends on the nature of the data being examined and probing of the special features revealed by graphical analysis or by visual examination. We consider some examples.

2.1. Editing, adjusting and faking

Let us look at the following table which appears on page 74 of the book, "Epidemiology, Man and Disease" by J.P. Fox, C.E. Hall and L.R. Elveback.

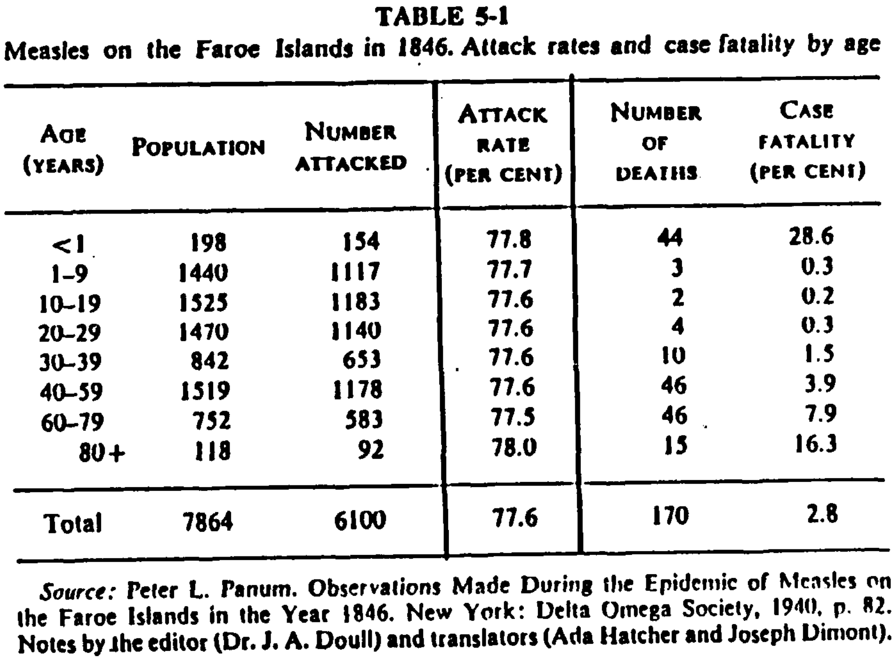


TABLE 5-1
Measles on the Faroe Islands in 1846. Attack rates and case fatality by age

AGE (YEARS)	POPULATION	NUMBER ATTACKED	ATTACK RATE (PER CENT)	NUMBER OF DEATHS	CASE FATALITY (PER CENT)
<1	198	154	77.8	44	28.6
1–9	1440	1117	77.7	3	0.3
10–19	1525	1183	77.6	2	0.2
20–29	1470	1140	77.6	4	0.3
30–39	842	653	77.6	10	1.5
40–59	1519	1178	77.6	46	3.9
60–79	752	583	77.5	46	7.9
80+	118	92	78.0	15	16.3
Total	7864	6100	77.6	170	2.8

Source: Peter L. Panum. Observations Made During the Epidemic of Measles on the Faroe Islands in the Year 1846. New York: Delta Omega Society, 1940, p. 82. Notes by the editor (Dr. J. A. Doull) and translators (Ada Hatcher and Joseph Dimont).

The authors conclude that "although the attack rates are high in all age groups, the fatality varied significantly, being highest under one year and then rising steadily for those over age thirty." Is this conclusion valid?

What is of interest to note in the table is the almost uniform attack rates of measles for all age groups (indicated by blocking). Could this occur by chance even if the true attack rate is common to all age groups? There is a strong suspicion that the number attacked in each age group was not observed but reconstructed from the known population size of that age group by multiplying it by the common overall attack rate of 6100/7864 = .776 and rounding off to the nearest integer. Thus the figures 154 for age less than 1 and 92 for over 58 could have been obtained as follows:

$$198 \times .776 = 153.648 \sim 154; \quad 118 \times .776 = 91.568 \sim 92. \qquad (2.1.1)$$

Now, if we use these reconstructed numbers to calculate the attack rates we get the values

$$\frac{154}{198} = .7777 \ldots \sim .778; \ \frac{92}{118} = .7796 \sim .780 \qquad (2.1.2)$$

as reported by the authors and also explains why the attack rates differ slightly in the third decimal place. A reference to the original report by Panum in German revealed that the number attacked was not originally classified by age groups but the number attacked in each group was reconstructed in the manner explained in the equation (2.1.1) by the editor of the English translation assuming a uniform attack rate. The attack rates reported in the blocked column of the above table are not found in the table on page 82 of the English translation, which are probably computed by the authors of the book in the manner explained in (2.1.2). In view of this, the age specific fatality rates computed from the reconstructed values of the number attacked in each group and the consequent interpretation may not be valid. A statistician is often required to do detective type work! (The second entry in the blocked column should be 77.6!)

There are a number of papers starting with two early papers by Fisher (1936) and Haldane (1948) describing statistical methods for examining whether data are faked or not. Haldane has said, "Man is an orderly animal. He finds it very hard to imitate the disorder of nature". Generally, faking may be suspected when some regularity is observed in recorded data.

2.2. Measurement and recording errors, outliers

In any large scale investigation measurement and recording errors are inevitable. It is difficult to detect them unless they appear as highly discordant values not in line with the others. Care should be taken to see in designing an investigation that such errors are minimized. A built-in scrutiny program while measurements are being made in the field might alert the investigator when a reading looks suspicious and allow him to repeat the measurement and/or investigate whether or not the individual being measured belongs to the population under study.

The author had the opportunity to scrutinize vast amounts of data collected in anthropometric surveys. In one case, the entire data collected at great cost had to be rejected (see Mukherji, Rao and Trevor (1955)). When the number of recording and measurement errors in multivariate response data is not large, they could be detected by drawing histograms of individual measurements and ratios, plotting two dimensional charts for pairs of measurements, and computing the first four moments and measures of skewness and kurtosis, γ_1 and γ_2. These measures are specially sensitive to outliers.

TABLE 1. TEST STATISTICS γ_1 FOR SKEWNESS AND γ_2 FOR KURTOSIS FOR SOME ANTHROPOMETRIC MEASUREMENTS OF SIX MALE TRIBAL POPULATIONS
(From the Thesis of Dr. Urmila Pingle)

	male tribal populations									
Character	KOLAM		KOYA		MANNE		MARIA		RAJ GOND	
	γ_1	γ_2	γ_1	γ_2	γ_1	γ_2	γ_1	γ_2	γ_1	γ_2
H.B.	.16	−.62	.39	.37	1.02* .71*	4.54* .29	−.27	.48	−.30	.23
H.L.	−.14	−.06	.48	1.12	−.05	−.08	.05	−.09	−.32	.28
Bg.B.	.83* −.14	2.93* −.03	.17	.19	1.72* −.40	8.42* .27	−.17	−.63	−.12	−.61
T.F.L.	−.26	−.07	.44	.11	.66*	.32	−.05	−.10	−.04	−.24
U.A.L.	−.05	−.63	−1.95* −.30	6.88* .74	−.01	−.27	.13	.76	.14	−.40
L.A.L.	−2.17* .08	9.98* −.62	−.07	.59	.19	−.67	−.02	.28	−.06	−.67

The values in the second line for each character are calculated after omitting extreme observations.

Table 1 gives the values of γ_1 and γ_2 computed from the original data and after removing extreme values for a number of characteristics for different populations sampled. The sample size for each group was of the order of 50, and asterisks indicate significance at the 5% level. It is seen that the recomputed values of γ_1 and γ_2, after omitting one extreme value in each case, are in conformity with others.

Normal probability plots are recommended for testing departures from normality and detecting outliers. Some of these including the method of fractile graphical analysis with independent subsamples are described in Section 9.2 of Rao, Mitra, Matthai and Ramamurthy (1973). An alternative method which has other potentialities is as follows. Let $\underset{\sim}{Y}' = (y_1,\ldots,y_n)$ be the vector of order statistics in a sample of size n. Further let

$$\underset{\sim}{c}' = (c_1,\ldots,c_n) = E(\underset{\sim}{Y}'),\ V = E(\underset{\sim}{Y}\underset{\sim}{Y}')$$

when the parent population from which the sample is drawn is N(0,1). Then

$$E(\underset{\sim}{Y}) = \mu\underset{\sim}{c}_0 + \sigma\underset{\sim}{c},\quad \underset{\sim}{c}_0' = (1,\ldots,1)$$

and the graph of (y_i,c_i), $i = 1,\ldots,n$, should be close to a straight line. The graph may show departures of various kinds. There may be extreme values which do not conform to the straight line trend exhibited by the bulk of the data. In such a case, we may omit the extreme values below and above, consider the linear model

$$y_i = \mu + \sigma c_i + \epsilon_i,\ i = r,r+1,\ldots,r+s;\ \mathrm{cov}(\epsilon_i\epsilon_j) = \sigma^2 V_{ij},\ i,\ j = r,\ldots,r+s \tag{2.2.1}$$

and estimate μ and σ by using the Gauss-Markoff theorem with a given variance-covariance matrix for the errors. There may be other types of departures indicating a non-linear relationship between y_i and c_i. In such a case we may write the extended model

$$\underset{\sim}{Y} = \alpha_1\underset{\sim}{b}_1 + \alpha_2\underset{\sim}{b}_2 + \alpha_3\underset{\sim}{b}_3 + \alpha_4\underset{\sim}{b}_4 + \ldots + \underset{\sim}{\epsilon} \tag{2.2.2}$$

where, with $\underset{\sim}{c}_i' = (c_1^i,\ldots,c_n^i)$,

$$\underset{\sim}{b}_1 = \underset{\sim}{c}_0,\ \underset{\sim}{b}_2 = \underset{\sim}{c}_1,\ \underset{\sim}{b}_3 = \underset{\sim}{c}_2 - \frac{\underset{\sim}{c}_2'V^{-1}\underset{\sim}{c}_0}{\underset{\sim}{c}_0V^{-1}\underset{\sim}{c}_0},\ \underset{\sim}{b}_4 = \underset{\sim}{c}_3 - \frac{\underset{\sim}{c}_3'V^{-1}\underset{\sim}{c}_1}{\underset{\sim}{c}_1'V^{-1}\underset{\sim}{c}_1},\ldots$$

and obtain estimates of the α-coefficients

$$\hat{\alpha}_j = \underset{\sim}{b}_j'V^{-1}\underset{\sim}{Y}/\underset{\sim}{b}_j'V^{-1}\underset{\sim}{b}_j,\ V(\hat{\alpha}_j) = \sigma^2/\underset{\sim}{b}_j'V^{-1}\underset{\sim}{b}_j$$

by the Gauss-Markoff theorem. If we denote $\hat{\sigma}^2 = \Sigma(y_i-\bar{y})^2/(n-1)$, then

$$T_2 = W^{\frac{1}{2}} = \hat{\alpha}_2(\underset{\sim}{b}_2'V^{-1}\underset{\sim}{b}_2)/\hat{\sigma}(\underset{\sim}{b}_2'V^{-2}\underset{\sim}{b}_2)^{\frac{1}{2}},\ T_3 = \hat{\alpha}_3(\underset{\sim}{b}_3'V^{-1}\underset{\sim}{b}_3)^{\frac{1}{2}}/\hat{\sigma},\ T_4 = \hat{\alpha}_4(\underset{\sim}{b}_4'V^{-1}\underset{\sim}{b}_4)^{\frac{1}{2}}/\hat{\sigma}$$

provide test statistics for judging depatures from normality, of which T_2 can be recognized as the Shapiro-Wilk (1965) W statistic for normality test. High values of T_3 and T_4 would indicate asymmetry and non-normal kurtosis respectively. The reader is referred to Puri and Rao (1976) for further details on the use of the statistics T_2,T_3 and T_4 in detecting departures from normality.

2.3. Too much fuss about outliers?

In section 2.2, we have discussed about the appropriate methodology when outliers are discovered. But omission of outliers or spurious observations may result in loss of information in some cases as the following example shows.

Suppose we have N observations from a population with mean μ and standard deviation (s.d.) σ giving a mean value $\bar{x}$, and M spurious observations from a population with mean ν and s.d. σ giving a mean value $\bar{y}$. Let us ignore the fact that $\bar{y}$ arises from contaminating observations and estimate μ by $\hat{\mu} = (N\bar{x}+M\bar{y})/(N+M)$. Then denoting $\nu - \mu = \delta\sigma$,

$$E(\hat{\mu}-\mu)^2 = \frac{\sigma^2}{N+M}\left(1+\frac{M^2\delta^2}{N+M}\right) < V(\bar{x}) = \frac{\sigma^2}{N}$$

if $\delta^2 < M^{-1} + N^{-1}$ which is always true when $\delta \leq 1$ amd $M = 1$ whatever N may be. Thus under the mean squared error criterion, which is popular among statisticians, it pays to include a spurious observation from a population whose mean may differ by as much as one standard deviation from the parameter under estimation! Such an improvement can be of considerable magnitude in small samples.

2.4. Non-response, a skeletal example

Missing values in a data set pose serious problems. It often happens that the missing values have special characteristics and ignoring them may lead to biased results. For instance, if we have a sample of skeletal material dug out from graves, most of the skulls would be in a broken condition and only a subset of the possible measurements on a skull could be taken on each specimen. On a skull that is preserved intact, the four characteristics, C(cranial capacity), L(length), B(breadth) and H(height), could be measured. On others, either none, or only B, or B and H, or L, B and H could be measured depending on the extent of breakage. The problem is to estimate the mean values and variances and covariances of C,L,B and H for the original population of skulls, some of which are well preserved and others broken. It is the usual practice, as found in published papers on skeletal studies, to estimate the mean value and variance for each characteristic say, C, from the available measurements on C alone, and the correlation between two characteristics say, C and L, from the available measurements on C and L. An alternative method, which is reported even in a recent paper, was to assume a joint density for the distribution of C,L,B and H, derive the marginal density for each subset of the characteristics, write down the likelihood for each specimen using the relevant marginal density and estimate the unknown parameters by the method of maximum likelihood (m.l.) from the joint likelihood of all the specimens. Such a procedure assumes that the sample of skulls providing the measurements on any specified subset of C,L,B and H constitutes a random sample from the original population of skulls. This is not generally true. For instance Rao and Shaw (1948) found that the skulls which are well preserved (intact) have on the whole smaller measurements than those that are broken, which shows that the well preserved skulls in the sample are on the whole smaller in size and could not be considered as a random sample from the original population of skulls. What is our strategy in such a case?

We may assume that at least the sample of skulls providing the measurement B only constitutes a random sample from the original population. The likelihood for such a specimen is the marginal probability density of B. Then for a specimen providing the measurements B and H, we consider the conditional probability density of H given B as its likelihood. Similarly we consider the conditional likelihoods of L given B and H, and C given L, B and H, for the specimens with measurements B,H,L and B,H,L,C respecitvely. We can then estimate the unknown parameters based on the product of the likelihoods and conditional likelihoods appropriate to the observed specimens. If the joint distribution of B,H,L and C is multivariate normal with the mean vector and the covariance matrix as

$$\mu = \begin{pmatrix} \mu_B \\ \mu_H \\ \mu_L \\ \mu_C \end{pmatrix} \qquad \text{and} \qquad \Sigma = \begin{pmatrix} \sigma_{BB} & \cdots & \sigma_{BC} \\ . & \cdots & . \\ \sigma_{CB} & \cdots & \sigma_{CC} \end{pmatrix}$$

then the method of estimation is as follows. Let

$$\bar{b}^{(1)},\ s_{bb}^{(1)}$$

$$\bar{b}^{(2)},\ \bar{h}^{(2)},\ s_{bb}^{(2)},\ s_{bh}^{(2)},\ s_{hh}^{(2)}$$

$$\bar{b}^{(3)},\ \bar{h}^{(3)},\ \bar{\ell}^{(3)},\ s_{bb}^{(3)},\ s_{bh}^{(3)},\ldots,s_{\ell\ell}^{(3)}$$

$$\bar{b}^{(4)},\ \bar{h}^{(4)},\ \bar{\ell}^{(4)},\ \bar{c}^{(4)},\ s_{bb}^{(4)},\ s_{bh}^{(4)},\ldots,s_{\ell c}^{(4)},\ s_{cc}^{(4)}$$

be the means and covariances estimated from the samples providing measurements on B alone, on B, H alone, ... and so on. The trends in the observed means $\bar{b}^{(1)},\ldots,\bar{b}^{(4)}$; $\bar{h}^{(2)},\ldots,\bar{h}^{(4)}$; ... will show whether the different types of samples are homogeneous on the basis of which a decision could be taken to use the actual likelihood or the conditional likelihood for each type of specimen. In case it is decided to use conditional likelihoods in all cases the estimates of the unknown parameters could be obtained as follows.

Let the regression equations and residual mean squares computed from appropriate complete samples be

$$\begin{aligned} b &= a_0^{(1)} &&, \quad s_{bb} \\ h &= a_0^{(2)} + a_1^{(2)}b &&, \quad s_{h.b} \\ \ell &= a_0^{(3)} + a_1^{(3)}b + a_2^{(3)}h &&, \quad s_{\ell.h,b} \\ c &= a_0^{(4)} + a_1^{(4)}b + a_2^{(4)}h + a_3^{(4)}\ell, && \quad s_{c.h,b,\ell} \end{aligned}$$

and define the matrices

$$A = \begin{pmatrix} 1 & & & \\ -a_1^{(2)} & 1 & & \\ -a_1^{(3)} & -a_2^{(3)} & 1 & \\ -a_1^{(4)} & -a_2^{(4)} & -a_3^{(4)} & 1 \end{pmatrix}, \qquad \underset{\sim}{a} = \begin{pmatrix} a_0^{(1)} \\ a_0^{(2)} \\ a_0^{(3)} \\ a_0^{(4)} \end{pmatrix}$$

$$S = \begin{pmatrix} s_{bb} & & & \\ & s_{h.b} & & \\ & & s_{\ell.h,b} & \\ & & & s_{c.\ell,h,b} \end{pmatrix}.$$

Then the estimates of Σ and μ are

$$\hat{\Sigma} = A^{-1}S(A^{-1})', \quad \hat{\mu} = A^{-1}\underset{\sim}{a}.$$

2.5. Regression analysis

There is considerable literature on the subject of regression analysis covering methods of estimation, detection of outliers and influential observations, selection of independent variables, transformation of variables and non-linear regression. Any software program for regression currently available has provision for computing the regression coefficients by a robust procedure, plotting the residuals to detect outliers and influential observations and revising the estimates of regression coefficients after possibly omitting some observations. However, there are certain other methods of multivariate data analysis which might be of use in examining in greater depth the nature of the relationship between the dependent and independent variables. Let us represent the data in a regression problem by $(\underset{\sim}{Y}:X)$ where $\underset{\sim}{Y}$ is a n-vector of dependent variables and X is an n×p matrix of independent variables, all expressed as deviations from the corresponding averages. Further let $S = X'X$, $T = (Y:X)'(\underset{\sim}{Y}:X)$, $\underset{\sim}{h}_i = X_i'S^{-1}X_i$, where X_i' is the i-th row of X and r_i be the residual (observed minus fitted value) at the i-th data point. Then the following analyses are suggested.

(i) Understanding the configuration of the independent variables: A cluster analysis of the data set X considered as n points in R^p, with the norm of $x \in R^p$ as $(nx'S^{-1}x)^{\frac{1}{2}}$, would show whether there are any gaps in the configuration of X which are relevant for the interpretation of results.

(ii) Understanding the joint configuration of the independent and dependent variables: A cluster analysis of the data set $(\underset{\sim}{Y}:X)$ considered as n points in R^{p+1}, with the norm of $x \in R^{p+1}$ as $(nx'T^{-1}x)^{\frac{1}{2}}$, would show whether the data break up into several clusters requiring a different regression function for each cluster. Other norms could also be tried.

(iii) Generally the residuals r_i are plotted against the serial number i or the fitted values $\hat{y}_i$. It will be of help in interpretation of data to plot r_i against $\sqrt{1-h_i}$ as all the residuals in any column will then have the same standard error. Further, the graphs of actual residuals or jack-knife residuals (calculated by leaving one out) will be similar. Again the whole configuration of residuals has to be examined to get a complete story and not merely looking for extreme values.

2.6. Graphical techniques

Portraying data graphically certainly contributes toward a clearer and more penetrative understanding of data and often provides clues for choosing appropriate stochastic models for inferential data analysis. With the sophisticated computer graphic facilities now available, the statistician is able to look at many plots during the statistical analysis and thus interact with data in a more effective way. Gnanadesikan (1977) describes a variety of graphical representations of multiresponse data, to test for multivariate normality, to detect outliers and to determine clusters. More recently, other types of graphical representations such as correspondence analysis (Benzécri and Benzécri (1980)), and projection pursuit (Friedman and Tukey (1974)) have been introduced which are receiving wide applications. The possibility of plotting higher dimensional data in a lower dimensional space as an aid in cluster analysis was first demonstrated in Rao (1948).

A word of caution is necessary in interpreting such a graphical representation of higher dimensional data in a lower dimensional space (see Rao (1971)). There is bound to be some distortion of relationships between units (or individuals) in such a representation. Any cluster of units that emerges from the graph has to be re-examined on the basis of the actual inter-unit distances in the original space. It may so happen that all the units in the observed cluster are not close to each other and they have to be divided into further clusters on the basis of differences in dimensions not represented in the graph.

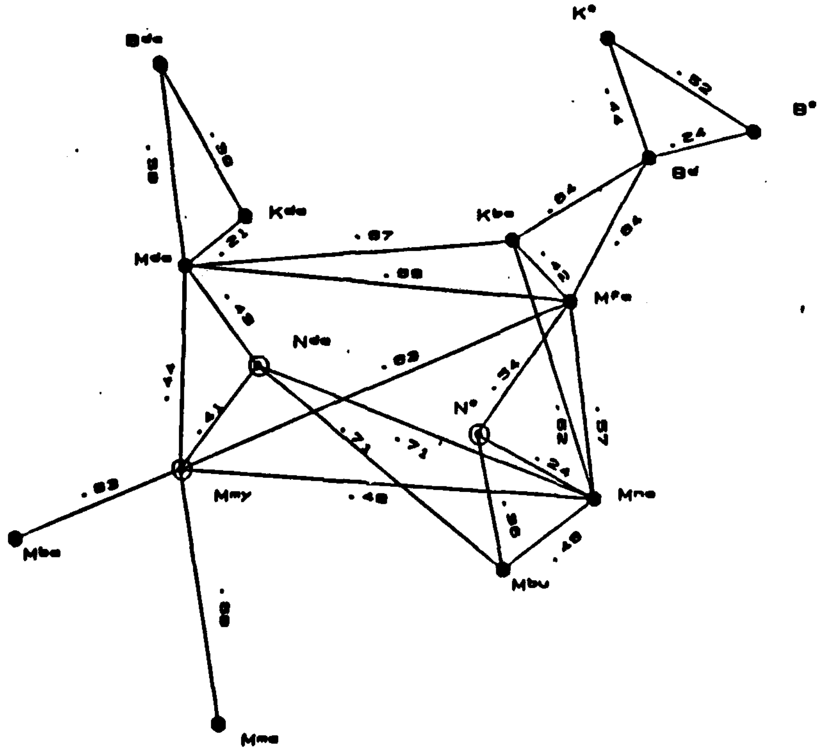


Figure 1. The vertices represent different castes studied in the Bengal Anthropometric Survey and the numbers represent Mahalanobis distances. (See Majumdar and Rao (1958)).

Since the fifties many methods of cluster analysis have been developed. Most of them end up with a dendrogram which provides distinct clusters of units at different levels of cluster distances. In practice distinct clusters rarely exist and the interrelationships between units are usually complicated. I have advocated listing of clusters of units, which may be overlapping, such that the units within a cluster have similarity coefficients less than a chosen threshold value. Such lists can be made at different threshold values. A listing of clusters at any threshold value can be represented as a graph with complete subgraphs (shown in Figure 1) where the vertices represent the units and all the vertices with distances less than a specified threshold value are connected. Such a figure has more information than a dendrogram. For instance the central positions occupied by the units M^{my}, N^{da} and N^{o} could not have been inferred from a dendrogram where they would have been classified with some clusters at some level of dissimilarity.

Listing of complete subgraphs at different threshold values can be time consuming if the number of units is large. The best strategy is first to determine broad clusters at a fairly high inter-cluster distance using a dendrogram and attempt a detailed study within each such cluster by the complete subgraph method.

3. Specification

Specification is the choice of a stochastic model in terms of which the observed data is analyzed, uncertainties in estimates of unknown parameters and tests of hypotheses are expressed and future observations are predicted. A chosen stochastic model, which may be called a working model, may not include the probability distribution (p.d.) which generated the observed data; this constitutes specification error. Then we have the problem of estimating the unknown parameters in a chosen model on the basis of given data and thus identifying a particular p.d. as close to or an estimate of the true p.d.; this process involves estimation error. Usually the specification and estimation errors balance each other so that a finer specification may not necessarily yield better results.

It should also be borne in mind that specification can depend on the purpose for which data analysis is undertaken. It is quite possible that different specifications for the same data have to be used for answering different questions.

How do we choose between competing models for a specified purpose? Several model selection criteria have been proposed (as the maximum likelihood principle does not provide a satisfactory answer) such as Akaike's information criterion, which are more appropriate in large samples where the specified purpose may not play a dominant role. However, in small samples a more appropriate method seems to be corss-validation as illustrated in the following Sections 3.1 and 3.2 with reference to a particular data set.

3.1. Cross-validation.

It is a technique by which a choice can be made between competing models by assessing the loss involved in using an estimated probability distribution. The idea is an old one which was used in testing models for weather prediction in the twenties. The data are subdivided into two parts: the first part is used to fit the model, and the second to validate it. It is only recently that the method has been modified, endowed with a suitable theory and successfully applied in many areas of research (Mosteller and Tukey (1968) and Stone (1974)).

An important application of cross-validation is in the selection of variables in multiple regression analysis. Let $E(Y) = X_{(s)}\beta_{(s)}$ be the linear model based on a subset (s) out of the p independent variables available. We compute the jack-knife residual

$$r_{i(s)} = y_i - X'_{i(s)}\hat{\beta}^{(i)}_{(s)} \tag{3.1.1}$$

where $\beta_{(s)}$ is estimated by omitting the i-th data point. Then the cross-validation error in prediction based on the specified subset of independent variables is

$$\text{CVE}(s) = n^{-1}\sum_{i=1}^{n} r^2_{i(s)}. \tag{3.1.2}$$

For different choices of subsets of the variables we compute (3.1.2) and choose that subset for which it is a minimum. Finally, for prediction purposes we estimate the regression coefficients based on the chosen subset of the variables using all the data points.

What is of interest in the above method is that we need not take the summation over all n data points as done in (3.1.2). If future prediction is needed in a specified region of the independent variables, we need only take the summation over those data points for which the observed independent variables are close to the specified region. The selected subset of independent variables may then depend on the specified region of the independent variables for future prediction.

3.2. A prediction problem

We shall use the cross-validation technique in deciding on a suitable rule for predicting the seventh measurement y_7 (the weight of a mouse at the seventh period) given the previous measurements $y_1,\dots,y_6$ (the weights taken at 6 previous periods). We have data on all the seven measurements for 13 mice on the basis of which we wish to determine the prediction formula (Rao (1987) gives the original measurements). The following formulas are tried.

1. Direct regression of y_7 on the subsets of previous measurements y_1-y_6, y_2-y_6, y_3-y_6, y_4-y_6, y_5-y_6 and y_6, using the least squares method of estimation.

2. Inverse regression method which predicts y_7 as

$$\tilde{y}_7 = \bar{y}_7(1-R^{-2}) + R^{-2}\hat{y}_7$$

where $\bar{y}_7$ is the sample average of y_7, R is the multiple correlation of y_7 on subsets of y_1-y_6, and $\hat{y}_7$ is the direct regression of y_7 on subsets of y_1-y_6.

3. A polynomial of a suitable degree is fitted to the first six or subsets of measurements of each mouse and the seventh value is predicted by extrapolation.

4. Regression of y_7 on the predicted value in (3) above. This provides some kind of calibration of the value in (3).

5. Empirical Bayes predictor using polynomial regression in (3). The polynomial regression coefficients are considered as random variables with a prior distribution function which can be estimated from the data as shown in Rao (1975).

6. Principal component regression using the first few principal components. The first six measurements are replaced by their principal components and the regression of y_7 on the first few principal components is computed.

7. Factor analytic regression using the first few factors. A model of the form

$$y_i = a_{i1}f_1 + \dots + a_{ik}f_k + \epsilon_i,\ V(\epsilon_i) = \sigma^2,\ i = 1,\dots,7,$$

is fitted to the entire data using all the seven variables, where a_{ij} are factor loadings and f_i are factors. For prediction in a future case, the factor values are estimated on the basis of the first six observations, considering the estimated a_{ij} as fixed, and substituted in the seventh equation to predict y_7.

The CVE's in all the cases are reported in Table 3. It is interesting to note that for purposes of prediction, much of the information is contained in the previous one or two measurements and modelling for the growth curve over the whole observed period for extrapolation introduces more noise in prediction. This shows that in problems of prediction greater attention should be given to obtaining a good fit in the neighborhood of the predictor values at which future prediction is required.

TABLE 3. CVE of different predictors for y_7 (Mice data, n=13)

Previous values used	Regression of y_7			Polynomial regression		
	direct	indirect	degree	indiv. regression predictor	calib-rated value	empirical Bayes
y_1-y_6	.095	.103	5	7.472	.252	-
			4	.600	.235	.375
			3	.175	.093	.139
			2	.104	.037	.087
			1	.206	.035	.194
y_2-y_6	.079	.081	4	2.405	.235	-
			3	.241	.141	.174
			2	.095	.040	.075
			1	.158	.035	.143
y_3-y_6	.047	.048	3	.757	.192	-
			2	.096	.052	.069
			1	.111	.034	.097
y_4-y_6	.037	.040	2	.229	.094	-
			1	.066	.034	.054
y_5-y_6	.031	.034	1	.055	.033	-
y_6	.027	.028	-	-	-	-

Principal component regression using all the measurements y_1-y_6 and first k principal components; .038 for k=1, .048 for k=2. Factor analytic regression using all the measurements y_1-y_7 and first k factors: .038 for k=1, .062 for k=2.

4. Inferential data analysis and some closing remarks

Inferential data analysis refers to the statistical methodology, based on a specified underlying stochastic model, for estimating unknown parameters, testing of hypotheses, prediction of future observations, making decisions etc. The choice of a model may depend on the specific information we are seeking from data. It may not necessarily be the one which explains the whole observed data as we saw in the problem of prediction (Section 3.2) using growth data.

Data analysis for answering specific questions raised by customers is not the only task of a statistician. A wider analysis for understanding the nature of given data would be of use in looking for other interesting questions which can be answered with available data, in raising new questions and in planning future investigations.

It is also a good practice to analyse given data under different alternative stochastic models and to examine differences in conclusions that emerge. Such a procedure may be more illuminating than seeking for robust inference procedures to safeguard against possible alternative stochastic models. The possibility of using different models for the same data to answer different questions should be explored.

Inferential data analysis should be of an interactive type as new features of the data may emerge during the analysis under a specified model requiring a change in the analysis originally contemplated.

Simulation studies to assess the performance of certain procedures and bootstrap and jack-knife techniques for estimating variances of estimators (Efron (1979)) under complicated data structures, which depend on the heavy use of computers, have given additional dimensions to data analysis, although some caution is needed in interpreting the results of such analyses.

There is the usual dictum in inferential data analysis that once the validity of a model is assured, there is an optium way of analysing the data such as the use of $\bar{x}$ as an estimate of the mean of a normal population based on a given sample, or of the mean of a finite population based on a random sample without replacement. As an example of the latter case, suppose that the problem is that of estimating the average yield of trees planted in a row by taking a sample of size 3. Our prescription says that if x_1, x_2, x_3 are the observed yields on three randomly chosen trees, then a good estimate is $\bar{x} = (x_1+x_2+x_3)/3$. However, if after drawing the sample we find that two of the trees chosen are next to each other with the corresponding yields, say x_1 and x_2, then we may be better off in giving the alternative estimator $\bar{x}' = (y+x_3)/2$ where $y = (x_1+x_2)/2$. It may be seen that if the yields of consecutive trees are highly correlated, then the variance of $\bar{x}'$ is less than that of $\bar{x}$ in samples where at least two consecutive trees are chosen. Such strategies as using different methods for different configurations of the sample under the same stochastic model should be explored. Another example of this type was considered in (2.1.1).

As statisticians, we are asked to advise on the appropriate statistical methodology (or software package program) for a certain data set. Our answer should be: statistical treatment cannot be prescribed over the phone or bought over the counter. The data has to be subjected to certain diagnostic tests and immediate complications taken care of, and then a course of treatment is prescribed and the progress is continuously monitored to decide on any changes in treatment if needed.

Summary

The purpose of statistical analysis is "to extract all the information from observed data". The recorded data may have some defects such as recording errors and outliers and the first task of a statistician is to scrutinize or cross-examine the data for possible defects and understand its special features. The next step is the specification of a suitable stochastic model for the data using prior information and cross-validation techniques. On the basis of a chosen model, inferential analysis is made, which comprises of estimation of unknown parameters, tests of hypotheses, prediction of future observations and decision making. Examining data under different possible models is suggested as more informative than using robust procedures to safeguard against possible alternative models. Finally data analysis must also provide information for raising new questions and planning future investigations. Some aspects of data analysis as outlined above are illustrated through examples.

Résumé

Le but de l'analyse statistique est "d`extraire toute l`information contenue dans des données observées. Les données enregistrées peuvent contenir des erreurs telles qu'erreurs de transcription et valeurs aberrantes. La première tâche du statisticien est de scrutiner et d'examiner les données afin de détecter ces erreurs et afin de comprendre les caractèristiques spéciales de ces données. L` étape suivante est la spécification d'un modèle stochastique adèquat pour les données en utilisant de l'information connue et des techniques de cross-validation. A partir d'un modèle choisi une analyse d'inférence est faite ce qui inclus estimation de paramètres inconnus, tests d hypothèses, prédiction de futures observations et prise décision. Il est proposé, parce que plus informatif, d`examiner les données sous différents modèles au lieu d`utiliser des procédures robustes pour se protéger de possibles modèles alternatifs. Finalement l`analyse de données doit aussi procurer de l`information afin de soulever de nouvelles questions et de planifier de futures investigations. Certains des aspects de l`analyse de données qui ont été mentionnés ci-dessus sont illustrés à l`aide d`exemples.

BIBLIOGRAPHY

1. Benzécri, J.P. and Benzécri, F. (1980). L'Analyse des Correspondesces: Exposé Elémentaire, Dunod, Paris.

2. Efron, B. (1979). Bootstrap methods: another look at jack-knife. Ann. statist. 7, 1-26.

3. Fisher, R.A. (1922). On the mathematical foundations of theroetical statistics. Philos. Trans. Roy. Soc. 222, 309-368.

4. Fisher, R.A. (1934). The effect of methods of ascertainment upon estimation of frequencies. Ann. Eugen. 6, 13-25.

5. Fisher, R.A. (1936). Has Mendel's work been rediscovered? Annals of Science 1, 115-137.

6. Freidman, J.H. and Tukey, J.W. (1974). A projection pursuit algorithm for exploratory data analysis. IEEE Transactions on Computers C-23, 881-889.

7. Gnandesikan, R. (1977). Statistical Data Analysis of Multivariate Observations, Wiley, New York.

8. Hacking, Ian (1984). Trial by number. Science 84, 69-70.

9. Haldane, J.B.S. (1948). The faking of genetic results. Eureka 6, 21-28.

10. Mahalanobis, P.C. (1931). Revision of Risley's anthropometric data relating to the tribes and castes of Bengal. Sankhyā 1, 76-105.

11. Mahalanobis, P.C. (1944). On large scale sample surveys. Philos. Trans. Roy. Soc. London, Series B, 231, 329-451.

12. Majumdar, D.N. and Rao, C. Radhakrishna (1958). Bengal anthropometric survey, 1945: A statistical study. Sankhyā, 19, 201-408.

13. Mosteller, F. and Tukey, J.W. (1968). Data analysis including statistics. In Handbook of Social Psychology, Vol. 2, (Eds. G. Lindzey and E. Aronson), Addison-Wesley.

14. Mukherji, R.K., Rao, C.R. and Trevor, J.C. (1955). The Ancient Inhabitants of Jebel Moya. Cambridge University Press.

15. Neyman, J. and Pearson, E.S. (1966). Joint Statistical Papers by J. Neyman and E.S. Pearson, Univ. of California Press, Berkeley.

16. Pearson, K. (1948). Karl Pearson's Early Statistical Papers, Cambridge University Press.

17. Pitman, E.J. G. (1937). Significance tests which may be applied to samples from any population. J. Roy. Statist. Soc. Ser. B, 4, 119-130.

18. Puri, M.L. and Rao, C. Radhakrishna (1976). Augmenting Shapiro-Wilk test for normality. Contributions to Applied Statistics, Birkhauser (Grossehaus), Berlin, 129-139.

19. Rao, C. Radhakrishna (1948). The utilization of multiple measurements in problems of biological classification. J. R. Statist. Soc. B 10, 159-203.

20. Rao, C. Radhakrishna and Shaw, D.C. (1948). On a formula for the prediction of cranial capacity. Biometrics 4, 247-258.

21. Rao, C. Radhakrishna (1971). Taxonomy is anthropology. In Mathematics in Archaelogical and Historical Sciences, Edin. Univ. Press, 329-358.

22. Rao, C. Radhakrishna (1975). Simultaneous estimation of parameters in different linear models and applications to biometric problems. Biometrics 31, 545-554.

23. Rao, C. Radhakrishna (1987). Prediction of future observations in growth curve models. Statistical Sciences (in press).

24. Rao, C. Radhakrishna, Matthai,A., Mitra, S.K. and Ramamurthy, G. (1973). Formulae and Tables for Statistical Work, Stat. Publishing Soc., Calcutta.

25. Shapiro, S.S. and Wilk, M.B. (1965). An analysis of variance test for normality. Biometrika 52, 591-611.

26. Shewhart, W.A. (1931). Economic Control of Quality of Manufactured Product, D. Van Nostrand, New York.

27. Stone, M. (1974). Cross-validation choice and assessment of statistical predictions. J. R. Statist. Soc. B 36, 111-133.

28. Tukey, J. (1962). The future of data analysis. Ann. Math. Statist. 33, 1-67.

29. Tukey, J. (1977). Exploratory Data Analysis (EDA), Addison Wesley.

30. Wald, A. (1950). Statistical Decision Functions, Wiley, New York.

END

12-87

DTIC